MOLTO
WP5: Statistical and Robust Translation

Lluís Màrquez
TALP Reseach Center

Software Department, UPC
MOLTO Kickoff meeting
Barcelona, March 9–11, 2010

WP5: Statistical and Robust Translation

Summary

- Goal: to develop hybrid MT methods that complete the GF-based methods of WP3 by extending their coverage in unconstrained text translation
- Scenario: *Patents* domain (WP7); $\geq$3 languages
- Work Plan:
  1. Probabilistic extension of a GF domain grammar
  2. Adapt base SMT systems to the *Patents* domain
  3. Develop and test hybrid GF-SMT translation methods

WP5: Effort

Start: M7; **End:** M30

Type	RTD				
Participant	UGOT	UHEL	UPC	Mxw	Ontotext
PMs	9	3	32	6	0

Total PMs: 50.0

UPC $\longrightarrow$ package leader, SMT technology, hybrid models, corpora processing

UGOT $\longrightarrow$ probabilistic extension of GF, synthetic corpora for SMT

Mxw $\longrightarrow$ relation with WP7, corpora provider

UHEL $\longrightarrow$ usability of the combined system

WP5: Description of work

1. Statistical GF domain grammar for *Patents*:

   1.1 Probabilistic extension

   - to cope with ambiguity,
   - provide cofidence–rated translations (rankings),
   - analyze partial phrases and to provide several partial translations,
   - increase robustness

   1.2 Automatic learning of grammars

WP5: Description of work

2. Train and adapt a SMT system to the *Patents* domain
   - Use large out-of-domain corpora available to create the base SMT system
   - Use small parallel corpora from WP7 for adaptation
   - Explore the usage of synthetic corpora generated by GF

WP5: Description of work

3. Develop hybrid approaches by combining GF and SMT
   - Cascade of independent MT systems (baseline)
   - Hard integration: fixing secure GF partial output in a probabilistic decoding
   - Soft integration: GF scored partial output integrated as new features in SMT decoding (either phrase or syntax–based)

WP5: Deliverables

D5.1 Description of the final collection of corpora (**M18**)

D5.2 Description and evaluation of the combination prototypes (**M24**)

D5.3 WP5 final report: statistical and robust MT (**M30**)

WP5: Milestones

MS5 First prototypes of the cascade-based combination models (**M18**)

MS7 First prototypes of hybrid combination models (**M24**)

MS8 Translation tool complete (**M30**)

WP5: Expected First Year Results

- From month 7 to 18:
  1. Compilation and annotation of corpora from the *Patents* domain
  2. Training and adaptation of the base SMT systems
  3. Statistical extension of the GF grammar
  4. Evaluation and comparison of GF and SMT systems in real domain data
  5. First experiments with the combination approaches (baseline plus hard integration)

MOLTO
WP5: Statistical and Robust Translation

Lluís Màrquez
TALP Reseach Center

Software Department, UPC
MOLTO Kickoff meeting
Barcelona, March 9–11, 2010